Personality classification based on Twitter text using Naive Bayes, KNN and SVM

Conference Paper · November 2015

DOI: 10.1109/ICODSE.2015.7436992

2 authors:

Bayu Yudha Pratama
University of Indonesia
1 PUBLICATION **71** CITATIONS
SEE PROFILE

Riyanarto Sarno
Institut Teknologi Sepuluh Nopember
270 PUBLICATIONS **1,683** CITATIONS
SEE PROFILE

Personality Classification Based on Twitter Text Using Naive Bayes, KNN and SVM

Bayu Yudha Pratama
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, 60111, Indonesia
bayu11@mhs.if.its.ac.id

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, 60111, Indonesia
riyanarto@if.its.ac.id

***Abstract*—Personality is a fundamental basis of human behavior. Personality affects the interaction and preferences of an individual. People are required to take a personality test to find out their personality. Social media is a place where users express themselves to the world. Posts made by users of social media can be analyzed to obtain their personal information. This experiment uses text classification to predict personality based on text written by Twitter users. The languages used are English and Indonesian. Classification methods implemented are Naive Bayes, K-Nearest Neighbors and Support Vector Machine. Testing results showed Naive Bayes slightly outperformed the other methods.**

Keywords— machine learning; personality identification; social media; text classification

I. INTRODUCTION

Personality is a combination of characteristic and behavior of an individual in dealing with various situations. Personality can influence a person's choice in various things such as websites, book, music and films [1]. Additionally, personality also affects the interaction with other people and environment. Personality can be used as assessment in employee recruitment, career counseling, relationship counseling and health counseling.

A person must take various personality tests to find out their personality. Personality tests can be self-descriptive report, interview or observation conducted by psychologists. These traditional methods are costly and less practical. Lately there has been a personality test in the form of an online questionnaire via website [2]. While this can be quite practical, user still has to take action in answering various questions. A recent study shows that personality traits can be automatically obtained from the text that they wrote [3]. The choice of most frequently used words can describe the personality of that particular person.

Social media is a place where users represent themselves to the world. Social media account is private and personal so it can reflect their personal lives. Activities in social media such as posting, commenting and updating status can reveal personal information. Text left by users can be analyzed to obtain information, in this case is personality of the user.

II. PREVIOUS RESEARCH

Big Five personality traits are five domains or dimensions of personality that are used to describe human personality [4]. The five factors are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Big Five is the most researched personality model in Psychology. Big Five are found in a variety of ages, cultures, and locations. It also consistently found in interviews, self-descriptions, and observations.

An experiment conducted to predict personality from features found in Facebook [5]. Features used are linguistic English words based on categories in LIWC (Linguistic Inquiry and Word Count) program [6], structural network, activity record and other personal information. Analysis performed using WEKA (Waikato Environment for Knowledge Analysis) program [7] with two built-in algorithms, M5 Rules and Gaussian Processes.

Naive Bayes method used to determine the user's personality from written [8]. User write self-descriptive text that will be used to find out their personality and then match them to find a partner on online dating site. The language used is Indonesian. The personality model used is the Four Temperaments. Four factors are Sanguine, Choleric, Melancholic, and Phlegmatic.

Emotion detection can also be predicted from text [9]. K-Nearest Neighbors used to determine emotion in documents. This experiment used Indonesian language. Text document is in the form of online news articles. Basic emotions include joy, anger, fear, sadness, disgust, and surprise.

Previous research in Big Five personality from social media Facebook based on text are in [10] and [11]. The corpus used is MyPersonality dataset [12]. These experiments use WEKA in analysis and predict the result. Various built-in algorithms used with accuracies in the range of 52%-61%.

In this experiment, we will try to improve accuracy from previous research. Furthermore, we discuss how to build a working system to predict personality from text written by the users of social media Twitter.

III. METHODOLOGY

Fig. 1 shows the overview diagram of this research. The system will retrieve a collection of tweets from users. Text from user then preprocessed into vector data. Classification process will classify user's text into a labeled dataset. The results are predictions for each Big Five traits, primary personality characteristics and secondary personality characteristics which obtained from the combination between two traits. System developed is a web application. The programming language used is Python with library Scikit-Learn [13].

978-1-4673-8430-8/15/$31.00 ©2015 IEEE

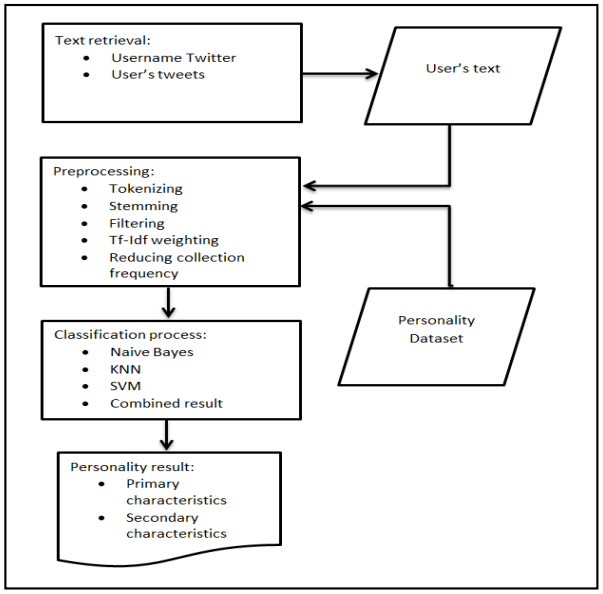

Fig. 1. Research overview diagram

A. Data Collection

This experiment uses MyPersonality dataset. MyPersonality Project is a Facebook application used to predict personality based on an online questionnaire. Dataset consists of 10.000 status updates from 250 users, which already labeled into Big Five personality dimensions. Original dataset then slightly modified. All posts from single user ID appended into one long string which considered single document. Final dataset is in the form of 250 documents from 250 users.

Indonesian text classification using the same dataset with the entire contents translated into Indonesian language as there is unavailability of Indonesian personality dataset. The assumption used is the meaning to every word remains accurate, despite being translated into another language. This method has limitations, as there is the possibility of mistranslation due the following reasons: ambiguous word, no equivalent word in Indonesia language or words that have different meanings in different contexts.

User text is taken from collection tweets of a Twitter user. System will take last 1.000 texts in the form of tweets (post made directly by the user) and re-tweets (re-posting someone else's text). Collection of tweets from users is also made into a single document/one long string.

B. Preprocessing Text

In text classification, text data will be represented in vector space model [14]. The steps in preprocessing text are as follows:

- Tokenizing: change sentence into a collection of a single word.
- Stemming: returns a word into basic form (root word) by eliminating existing additive. Stemming algorithm used is Porter Stemmer for English language and Nazief-Andriani for Indonesian language [15].
- Filtering: eliminating stop words. Stop word is a common word that has little or no meaning, but required in the structure of grammatical language.
- Weighting: calculate Tf-Idf for each word with (1).

$$tfidf_t = f_{t,d} \times log \frac{N}{df_t} \quad (1)$$

$tfidf_t$ = weight of term t
$f_{t,d}$ = occurrences term t in document d
N = total document
df_t = number document contains term t

- Reducing collection frequency or total number instances in dataset. MyPersonality contained ± 10.000 unique words. In this experiment classification, number of feature/word is limited to 750 words that most frequently appear in the dataset. Restriction on the number of words is to reduce the load and processing time, increase effectiveness, and improve accuracy.

C. Classification Process

The personality classification case is multi-label classification. It means one person can have more than one personality trait or do not have dominant personality trait at all. Multi-label method used in this experiment is a binary relevance that is transforming each label in binary with independent assumption [16]. The solution to this problem is to create a classifier for each label and train classifier based on the data that has been transformed. Each classifier is a binary classifier which will give output if the test document is a member of the label or not.

Naive Bayes is a classification algorithm based on the application of Bayes theorem [17]. Multinomial Naive Bayes (MNB) is a variation of the Naive Bayes designed to solve the classification of text documents. MNB uses multinomial distribution with the number of occurrences of a word or the weight of the word as a classification feature. MNB equation is shown in (2).

$$P\ (X|c) = \log \frac{N_c}{N} + \sum_{i=1}^{n} \log \frac{t_i + \propto}{\sum_{i=1}^{n} t\ + \propto} \quad (2)$$

$P(X|c)$ = probability document X in class c
N_c = total documents in class c
N = total documents
t_i = weight term t
$\sum_{i=1}^{n} t$ = total weight term t in class c
$\propto$ = smoothing parameter

K-Nearest Neighbors (KNN) is a classification algorithm that uses a distance function between the train data to test data and the number of nearest neighbors to determine the classification results. Distance function used in this experiment is the cosine similarity. Cosine similarity is one of the functions that are widely used in the document classification to find similarity between some documents [17]. Scoring function of KNN shown in (3). Determining document class is done by voting on K nearest neighbor. The nearest neighbor is the K-document with the highest similarity value.

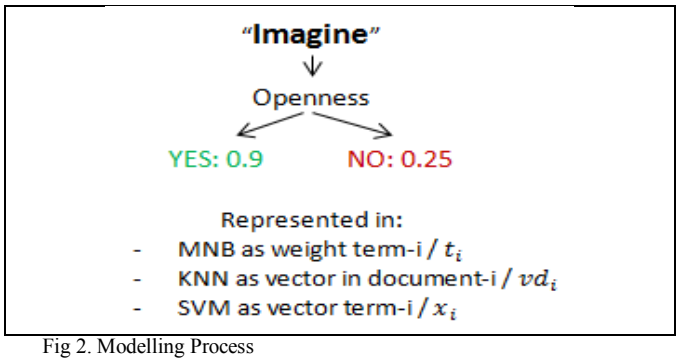

Fig 2. Modelling Process

$$score(c, d_1) = \sum_{d_2 \in S_k d_1} I_c(d_2) \cos\ (vd_1, vd_2) \quad (3)$$

$score(c, d_1)$ = scores of test document
d_1 = test document
d_2 = train document
vd_1 = vector test document
vd_2 = vector train document
I_c = 1 if d_2 is in class c, 0 otherwise
$S_k d_1$ = set of k nearest in test document

Support Vector Machine (SVM) is a supervised learning algorithm that analyzes the data and recognizing patterns used for classification [17]. SVM take the set of training data and marking it as part of a category then predicts whether the test document is a member of an existing class. SVM models represent the data as a point in space divided by a line/hyperplane. Optimum hyperplane search function shown in (4) subject to (5). Scoring in SVM to find the test document class use (6).

$$\frac{1}{2} w^T w + C \sum_i \xi_i \quad (4)$$

$$\{(x_i, y_i)\}, y_i(w^T x_i + b) \geq 1 - \xi_i \quad (5)$$

w = weight vector
C = loss function
ξ_i = slack variable/misclassification vector i
x_i = train vector i
y_i = class train vector i
b = bias value

$$f(x) = \text{sign}(w^T x + b) \quad (6)$$

$f(x)$ = score function
x = vector test document

Fig. 2 shows an example of modelling process used in a word. Each word has a probability value in each personality traits categories, which calculated in Tf-Idf weighting process. This value then used as weight or vector term in respective methods. In example, word "Imagine" frequently used by people with high Openness thus it has higher probability in "Yes" class. Each method then uses this value to make predictions.

Distribution of class member in the dataset is unbalanced. There is a personality class with number of people/member significantly greater than other classes. Tuning threshold for each classifier decision is needed in determining whether the test data is member in the label's personality (default is 0.5). The choice of the optimal threshold value is taken from the decision point that has highest F-Score in cross-validation testing. F-Score (9) is the mean value of True Positive Rate, shown in (7) and True Negative Rate (8). Table 1 and Table 2 shows the optimal threshold for every classifier of each method in English and Indonesia dataset.

$$TPR = \frac{TP}{TP+FP} \quad (7)$$

$$TNR = \frac{TN}{FP+TN} \quad (8)$$

$$\text{F-Score} = 2 \times \frac{TPR \times TNR}{TPR+TNR} \quad (9)$$

TPR = True Positive Rate/Sensitivity
TNR = True Negative Rate/Specificity
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative
F-Score = F-Score

The predicted result for each method can be different from one another. To avoid confusion in the conclusions, we propose a combined result which taken from three existing method's results. The combined result uses majority vote of the three methods. For example, if on a label there are two or more methods with the "Yes" prediction, then the final prediction is "Yes". Table 3 shows combination possibilities for combined method.

TABLE 1. Threshold Values for each Classifier in English Dataset

Label	MNB	KNN	SVM
AGR	0.58	0.56	0.25
CON	0.48	0.4	-0.15
EXT	0.34	0.45	-0.1
NEU	0.32	0.33	-0.1
OPN	0.77	0.74	0.25

AGR = Agreeableness
CON = Conscientiousness
EXT = Extraversion
NEU = Neuroticism
OPN = Openness

TABLE 2. Threshold Values for each Classifier in Indonesia Dataset

Label	MNB	KNN	SVM
AGR	0.59	0.59	0.15
CON	0.49	0.44	-0.1
EXT	0.35	0.41	-0.05
NEU	0.3	0.33	-0.05
OPN	0.76	0.74	0.3

TABLE 3. Combination Possibilities for Combined Method

MNB	KNN	SVM	Combined
Y	Y	Y	Y
Y	Y	N	Y
Y	N	Y	Y
Y	N	N	N
N	Y	Y	Y
N	Y	N	N
N	N	Y	N
N	N	T	N

IV. Result and Discussion

In this section, we report and discuss the performances of the classification algorithms on the personality traits recognition task. Testing was conducted using 10-fold cross-validations. Table 4 reports accuracy result for English dataset. Table 5 reports accuracy result for Indonesian dataset. In the cross-validation testing, MNB got the best accuracy in three methods tested with average accuracy 60%. SVM and KNN performed similarly. SVM method performs worse than MNB due to difficulties separating a class of a word as dataset are not quite accurate. KNN method also performs worse than MNB. The alleged cause of the low accuracy of the KNN method because of the difficulty in determining the optimal value of *K*. Total value of *K* is crucial because the KNN's probability result will be calculated from the *K* samples. This is different from MNB that uses pure probability calculations on existing features. Based on macro-averaged scores in 59%-60%, this experiment fails to improve accuracy, as it is only equal to the best score from previous research (61%).

The next scenario is respondent testing. This test intended to determine how this automatic personality prediction system fares against more traditional personality prediction. Right now one of most popular personality prediction test is online questionnaire based test. Therefore, we compare the system's result (Fig. 3) with the predicted result from the IPIP 50-Item Set of IPIP Big-Five Factor Markers questionnaires [18]. System will retrieve text data from users Twitter account and classified using three methods and the combined method. The users then complete questionnaire test and report the results. Respondents consisted of 40 people. Respondents chosen must have a Twitter account with a minimum number of 1.000 tweets. Classification language selection is based on user's primary language. The test results are shown in Table 6.

TABLE 4. Accuracy of English dataset

ACC	MNB	KNN	SVM
AGR	61.33 ± 1.89	57.38 ± 1.67	61.88 ± 1.97
CON	62.87 ± 1.54	63.00 ± 1.68	60.75 ± 1.16
EXT	57.75 ± 1.26	61.46 ± 2.11	59.99 ± 1.75
NEU	57.54 ± 1.9	52.25 ± 1.86	57.30 ± 1.29
OPN	63.67 ±1.54	64.92 ± 2.49	59.62 ± 1.54
AVERAGE	60.63 ± 1.64	59.8 ± 1.98	59.62 ± 1.54

TABLE 5. Accuracy of Indonesia dataset

ACC	MNB	KNN	SVM
AGR	63.21 ± 1.84	60.71 ± 2.48	62.95 ± 1.2
CON	58.67 ± 2.1	61.08 ± 3.19	60.35 ± 1.38
EXT	51.5 ± 2.02	52.87 ± 2.03	53.21 ± 2.17
NEU	54.63 ± 2.01	54.0 ± 1.83	54.2 ± 2.14
OPN	72.29 ± 2.13	62.83 ± 2.07	65.75 ± 1.69
AVERAGE	60.06 ± 2.02	58.30 ± 2.37	59.29 ± 1.76

TABLE 6. Accuracy of respondent testing

Method	ACC
MNB	63 %
KNN	60 %
SVM	61 %
Combined	65 %

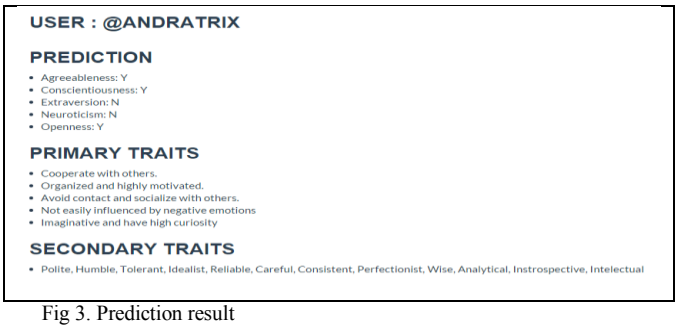

Fig 3. Prediction result

Combined method that is the final prediction of the application, got the best results on respondent testing with an accuracy of 65%. The combined method can produce better accuracy because there were improvements in the classification results. In the case where one of the methods fails at classifying, it will be covered by the two other methods if the predictions of other methods are correct. Overall accuracy is not quite good, but it can show that automatic text personality recognition can be alternative for questionnaire based test.

V. Conclusion

User's personality successfully predicted from text written on Twitter. In three methods used, Naive Bayes slightly outperformed the other methods with 60%. Experiment fails to improve accuracy from previous research. System has 65% accuracy compared to questionnaires based test.

Further improvement can be done by using more accurate dataset to improve the accuracy and using native Indonesian language for Indonesia classification (not translated). Future study also can include a semantic approach to consider the meaning of each word.

References

[1] I. Cantandir, I. Fernandez-Tobiaz, A. Belllogin, "Relating personality types with user preferences in multiple entertainment domains," EMPIRE 1st Workshop on Emotions and Personality in Personalized Services, 2013.

[2] L.R. Goldberg, J.A. Johnson, H.W. Eber, R. Hogan, M.C. Ashton, C.R. Cloninger, "The International personality item pool and the future of public domain personality measures," Journal of Research in Personality, 40(1), 84-96, 2006.

[3] F. Mairesse, M. Walker, M. Mehl, R. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," Journal of Artificial Intelligence Research (JAIR). 30(1), 457–500, 2007.

[4] P.T. Costa, R.R. McCrae, "Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI)," Psychological Assessment Resources, 1992.

[5] J. Golbeck, C. Robles, K.Turner, "Predicting personality with social media," In CHI'11 Extended Abstracts on Human Factors in Computing Systems, pp. 253-262, 2011

[6] J.W. Pennebaker, M.E. Francis, R.J. Booth. *Inquiry and Word Count: LIWC*, 2001.

[7] M. Hall, E. Frank, H. Holmes, P. Pfahringer, P. Reutemann, I.H. Witten, "The WEKA data mining software: an update," SIGKDD Explorations, Volume 11, Issue 1, 2009.

[8] N.M.A. Lestari, I.K.G.D. Putra, A.A.K.A. Cahyawan, "Personality types classification for Indonesian text in partners searching website using Naive Bayes methods," International Journal of Software and Informatics Issues, 2013.

[9] Arifin, K.E. Purnaha, “Classification of emotions in Indonesia text using K-NN method,” International Journal of Information and Electronics Engineering, Vol 2. No 6, 2012.

[10] A. Firoj, E.A. Speanov, G. Riccardi., “Personality traits recognition on social network – Facebook,” ICWSM Workshop on Personality Classification, 2013.

[11] F. Iacobelli, A. Cullota, “Too neurotic, not too friendly: structured personality classification on textual data”, ICWSM Workshop on Personality Classification, 2013.

[12] F. Celli, F. Pianesi, D. Stillwell, M. Kosinski, "Workshop on computational personality recognition (shared task)," In Proceedings of WCPR13, in conjunction with ICWSM-13, 2013.

[13] F. Pedregosa, et al., “Scikit-learn: machine learning in Python,” JMLR 12, pp 2825-2830, 2011.

[14] Ø. L. Garnes. *Feature Selection for Text Categorization*. Oslo, 2009.

[15] B.A.A. Nazief, M. Adriani, “Confix-stripping: approach to stemming algorithm for Bahasa Indonesia,” Internal publication, Faculty of Computer Science, Univ. of Indonesia, Depok, Jakarta, 1996.

[16] G. Tsoumakas, I. Katakis, “Multi-label classification: an overview,” International Journal of Data Warehousing & Mining 3 (3): 1–13, 2007.

[17] C.D. Manning, P. Raghavan, H. Schutze. *Introduction to Information Retrieva*l. Cambridge UP, 2008

[18] LR. Goldberg, "The Development of markers for the Big-Five Factor structure," Psychological Assessment 4.1: 26, 1992.